Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository

Catherine Plaisant, Jean-Daniel Fekete, Georges Grinstein

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS

Promoting Insight Based Evaluation of Visualizations: From Contest to Benchmark Repository

Catherine Plaisant
Human-Computer Interaction Lab.
University of Maryland
plaisant@cs.umd.edu

Jean-Daniel Fekete
INRIA Futurs/LRI
Université Paris-Sud
Jean-Daniel.Fekete@inria.fr

Georges Grinstein
Institute for Vis and Perception Research
University of Massachusetts Lowell
grinstein@cs.umd.edu

ABSTRACT

Information Visualization (InfoVis) is now an accepted and growing field but questions remain about the best uses for and the maturity of novel visualizations. Usability studies and controlled experiments are helpful but generalization is difficult. We believe that the systematic development of benchmarks will facilitate the comparison of techniques and help identify their strengths under different conditions. We were involved in the organization and management of three information visualization contests for the 2003, 2004 and 2005 IEEE Information Visualization Symposia, which requested teams to report on insights gained while exploring data. We give a summary of the state of the art of evaluation in information visualization, describe the three contests, summarize their results, discuss outcomes and lessons learned, and conjecture the future of visualization contests. All materials produced by the contests are archived in the Information Visualization Benchmark Repository.

General Terms

Visualization, information, competition, contest, benchmark, repository, measure, metrics.

1 Introduction

Information Visualization is now an accepted and growing field with numerous visualization components used in mainstream applications such as SPSS/SigmaPlot, SAS/GRAPH, and DataDesk, in commercial products such as Spotfire, Inxight, Tableau, HumanIT, and ILOG JViews, and in domain specific standalone applications such as interactive financial visualizations [SMo06] and election data maps [NYT06]. Nevertheless, questions remain about the potential uses of these novel techniques, their maturity and their limitations.

Plaisant reviewed evaluation challenges specific to information visualization and suggested initial actions [Pla04] such as refined evaluation methodologies, use of toolkits, dissemination of success stories, and the development of contests (Figure 1), benchmarks and repositories.

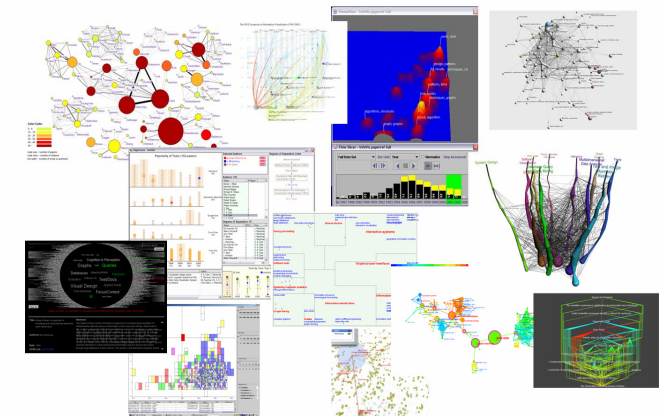

Figure 1: A collage of sample screens from the InfoVis 2004 contest submissions which illustrates the diversity of visualization methods used to address a task and highlights the difficulty of comparisons.

Empirical user studies are very helpful but take significant time and resources, and are sometimes found of limited use as they are conducted with ad-hoc data and tasks in controlled laboratory situations. Benchmarks facilitate the comparison of different techniques and encourage researchers to work on challenging problems. However to be convincing, the utility of new techniques needs to be demonstrated in a real setting, within a given application domain and a set of users. Contests attempt to create these surrogate situations that are representative of real world situations. They engage a competitive spirit and often produce results that help the community with comparisons of visualization tools applied to the same problem.

Competitions help push the forefront of a field quickly. TREC evaluations (the Text REtrieval Conference) [Voo00] exemplify the best of these in obtaining the participation of many corporate and academic research groups. For many it is simply the emotional aspect of winning or the excitement of the competition that compels them to participate.

A contest presents a problem that many will attempt to solve. If the problem is challenging and representative of a real world situation, then the contestants' solutions highlight what techniques are possible, and which

Digital Object Indentifier 10.1109/TVCG.2007.70412

solutions seem better to pursue. Participants can describe the insights they gained while using their tools. Insight can simply be defined as a non-trivial discovery about the data or, as a complex, deep, qualitative, unexpected, and relevant assertion [Nor06]. We believe that the Infovis 2003 contest was the first attempt to include the reporting of insights as an evaluation criterion [Inf03]. Often some solutions provide such good results that other participants are driven to compete in the next year's contests if the problem offered is similar. But for contests to have a long term impact, benchmark datasets, the associated tasks and the submitted materials need to be archived in a repository, safeguarding the baselines against which new techniques and approaches can be compared.

In this paper we briefly review the state of the art and the challenges for evaluation in information visualization. We also describe the first three information visualization contests, summarize their results, and discuss their effect.

2 Information Visualization Evaluation State of the Art

Information visualization systems can be very complex [Chen00] and require evaluation efforts targeted at different levels. One approach described in the evaluation section [Las05] of the Visual Analytics research agenda [Tho05] identifies three levels: the component level, the system level, and the work environment level (Figure 2).

The component level includes the individual algorithms, visual representations, interactive techniques and interface designs. Data analysis algorithms can often be evaluated with metrics that can be observed or computed (e.g. speed or accuracy), while other components require empirical user evaluation to determine their benefits. There have been demonstrations of faster task completion, reduced error rates or increased user satisfaction measured in laboratory settings using specific visualization components. For some techniques some scores can be computed to evaluate the potential quality of simple displays, e.g. [Mac86], but controlled experiments remain the workhorse of evaluation.

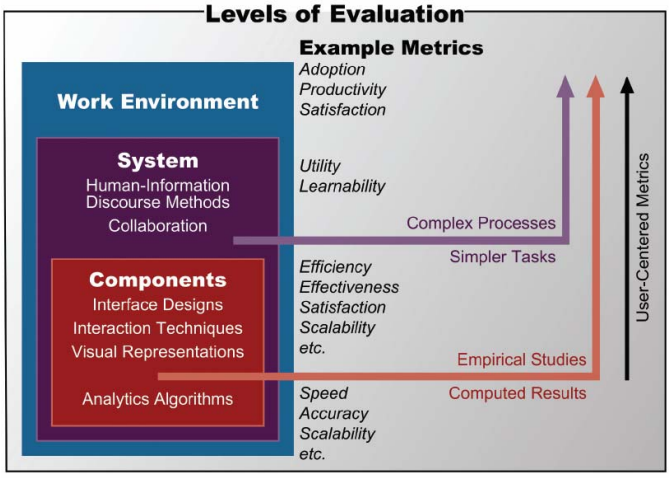

Figure 2: The three evaluation levels for Visual Analytics (Figure 6.1 in [Tho05])

The system level focuses on interfaces which combine and integrate multiple components and need to be evaluated by comparing them with technology currently used by target users (e.g. [Pla02]). Those evaluations also usually consist of controlled experiments that take place in the laboratory, using surrogate scenarios and short tasks. A more modern approach is to conduct insight-based evaluations where participants are given open-ended complex tasks and asked to report on the insights gained [Nor06, Sar04]. The discovery process is rarely an isolated, short-term process nor is it unique. Users may need to look at the same data with multiple tools over a longer period of time (days or months). This cannot be captured with controlled laboratory studies.

The third level is the work environment level where evaluation addresses issues influencing adoption. Case studies describe users in natural environment doing real tasks [Gon03, Tra00]. These are time consuming and may not be applicable to other domains but they can describe discoveries that take place over a long period. Shneiderman and Plaisant have proposed Multidimensional In-depth Long-term Case Studies (MILCS) as a way to study and evaluate creativity tools such as those in visual analytics and information visualization [Shn06].

Of course usability evaluation still remains the cornerstone of user-centered design and formative evaluation. It is not only of paramount importance for product engineering but also a powerful tool for researchers. It provides feedback on problems encountered by users and steers designers toward better designs at all three evaluation levels. The recent Beliv'06 workshop [BEL06] provided an excellent overview of the most recent work relevant to information visualization evaluation, including the

development of specific heuristics, metrics and task taxonomies.

3 Evaluation Programs and Efforts in Other Communities

Simple benchmark data sets abound. Some repositories simply make data sets available (e.g. the Council of European Social Science Archives [CES06]) while others offer tools to help promote research in specific domains (e.g. University of California Irvine Machine Learning Repository [MLR06]). Some repositories are clearly meant to promote evaluation (e.g. the Critical Assessment of Microarray Data Analysis or CAMDA conference [CAM06]).

A few research communities have a long history of success at promoting large scale evaluation programs and methodologies for measuring the effectiveness of information technology. For example TREC (the Text REtrieval Conference) [Voo00, TRE06] started in 1992 and provides datasets and scoring methods to evaluate information retrieval methods. The Message Understand Conference (MUC) series [Chi93] was started in 1987 for the purpose of qualitative evaluation of the state of the art in message understanding and has contributed to the development of improved linguistically-based systems. The speech recognition evaluation program [Pal00] has been creating speech corpora and benchmark tests primarily based on word error rates since 1988. The earlier work provides baselines against which researchers could demonstrate the effectiveness of recognition algorithms. Recent work has included the development of broadcast news transcription tasks to support topic detection and speech-to-text research.

Those sponsored evaluation programs are successful in part because clear computed metrics can be established to compare the results of tools with a trusted "ground truth", even if sometimes the ground truth is generated by humans as it is for TREC where human experts decide the "true" relevance of documents.

Other volunteer efforts have also been useful. The KDD Cup is an example of benchmarking where the knowledge discovery community agrees on scores that can be computed on new datasets created every year [Geh04].

Recently, the first Visual Analytics Science and Technology contest took place [VAS06; Gri06]. The VAST contest used a synthetic data set with embedded ground truth integrated into real text data in a careful manner [Whi06]. The scenario involved individuals performing some fraudulent activities and the discovery required analyzing and exploring the collection of multimedia documents (mostly text) and identifying the links between the individuals, places, and events. Ground truth was available about the "who, what, when, where and why" of suspicious activities facilitating evaluations as to how well the teams had analyzed the situation and permitting better estimates of the general utility of the tools. A pilot investigation looked at quantitative metrics using for example the number of subplots identified by each team and their relative complexity and subtlety, and found the ranking of the submissions to closely match the subjective assessments of the judges.

In the Human-Computer Interaction community contests are usually design competitions, one memorable exception being the 1997 CHI Browse-off [Mul97]. It brought together several visualization and browsing technologies for a live comparison, almost a competition, in which the tasks were generic information retrieval ones. In parallel with the development of benchmarks, specific evaluation methods and tools for emerging new technologies are being designed, for collaborative environments [Sch05], ubiquitous computing [Sch02], and intelligence analysis [Cow05].

Many other communities have evaluation programs. A complete survey of practices and lessons learned is needed and would help foster collaboration and coordination between evaluation efforts.

4 The Infovis Contests

4.1 The Contest Process

All three contests had a similar general organization. The contest was announced at the previous conference and the call for participation made public in February at the same time as the corresponding InfoVis Symposium call. The dataset and tasks were posted shortly thereafter. Participants had four months to prepare with a June deadline and had to submit a set of electronic documents consisting of

- a two page summary,
- a video illustrating the interactive techniques used,
- a detailed webpage describing the tool, how it was used to accomplish the tasks, and what insights were gained, and
- information about the team and tool provenance.

Judging was conducted during the summer primarily by the co-chairs and resulted in the selection of multiple first and second place entries. In October a session at the conference was dedicated to the contest, during which the chairs summarized the process and the lessons learned, and several teams presented their results. The

teams also demonstrated their systems during the poster session. Finally two page summaries were published in the adjunct proceedings and all materials archived in the Information Visualization Benchmark Repository [Bmr06].

4.2 Judging

The criteria for judging included

1) the quality of the data analysis (what interesting insights were found in the data),
2) the appropriateness of the visual representation for the data and the tasks, the usefulness and creativity of the interactivity, and the flexibility of the tool (how generic was the approach), and
3) the quality of the written case study (description of the strengths and weaknesses of the tool used).

Judges reviewed all the submissions, a very time consuming activity. A list of criteria was used, but numerical ratings were not necessarily compiled, (although attempted) because the number of entries was small enough that judges could easily group the entries in three categories and nominate the first and second place candidates. A number of conference calls and meetings allowed the chairs to decide on the number of winning entries and finalize the results.

Selecting the first place was generally easy as some entries stood out, but deciding on the cut-off for rejected entries was very difficult as many entries of lesser quality still had some interesting features we wanted to highlight and helped demonstrate the variety of solutions. We also wanted to encourage and reward participation.

Laskowski [Las05] provides a thorough review of the challenges of visual analytics evaluation many of which are common to information visualization evaluation. The contests placed the participants in a fairly realistic situation, giving them a fair but not arbitrarily large amount of time to analyze the data and prepare their answers. Although ideally we would have liked to evaluate the quality of answers computationally this was not possible. The problem was due to the fuzziness and variety of both the questions and answers to the contest. These included visual representations, collections of articles, new algorithms, or new visualization tools, each of whose correctness or evaluation may not be computable. This still forces human evaluation. TREC, for example, still uses human judging for determining the accuracy of the retrieval and such is the case for the IEEE InfoVis Contests we discuss.

Another difficulty for information visualization comes from the impact of the discovery process, an extremely interactive and personal activity. Whereas computational algorithms can be compared through the accuracy of their results, it is still not possible to accurately measure the results or impact of a single or multiple visualizations. We still do not have measures of perceptual information transfer. There is beginning research in measures of interestingness and other metrics related to visualization [Kei95, Gri02], but these are in their infancy and too simple to be applied to the current contests.

So we either propose simple tasks which yield precise results or specify more complex exploratory tasks and thus have much less predictable results. The latter is more interesting but makes the evaluation process difficult to plan for and forces real time evaluation criteria which sometimes end up reviewer dependent. One additional argument for simple tasks even if they are unrealistic is that a system which fails to achieve simple tasks would be a very limiting system and is likely not to support more complex or exploratory tasks. Our approach was to balance task simplicity and complexity to obtain a satisfying tradeoff.

We now review the 2003, 2004 and 2005 contest one by one in detail describing the data, tasks, judging, results and lessons learned.

4.3 The 1st contest – Infovis 2003

The first contest took place in 2003 [Inf03]. It focused on the analysis of tree structured data and in particular looking at the differences between similar trees.

4.3.1 The 2003 Data

There are hundreds of types of trees with varying characteristics. In an effort to represent this diversity in an accessible manner, contestants were provided three very different application examples with datasets in a simple XML format. These were

- Phylogenies - Small binary trees (60 leaf nodes) with a link length attribute. No node attributes except their names.
- File system and usage logs - Large trees (about 70,000 leaf nodes) with many attributes, both numerical and nominal. Changes between the two trees could be topological or attribute value changes. Data for four time periods was provided.
- Classifications - Very large trees (about 200,000 leaf nodes) with large fan-outs. Three node attributes, all nominal. Labeling, search and presenting results in context is important. We allowed teams to work on a

subset of the dataset (the "mammal" subtree) if they could not handle that many nodes.

4.3.2 The 2003 tasks

We provided general tasks (about 40 tasks in 11 categories) and tasks specific to the selected datasets. General tasks were low level tasks commonly encountered while analyzing any tree data: topological tasks (e.g., which branch has the largest fan-out?), attribute-based tasks (e.g., find nodes with high values of X), or comparison tasks (e.g., did any node or subtree "move"?).

The tasks specific to each dataset included more broad goal-setting tasks (e.g., for the phylogenies, what mapping between the two trees topologies could indicate co-evolution and, maybe, the points where the two proteins were not co-evolving?)

We made it clear that it was acceptable to submit partial results. For example, one could work only on some of the tasks. We also clarified that we were not looking for a detailed result list (e.g., a list of deleted nodes for the task "what nodes where deleted") but an illustration or demonstration of how the visualization helped find these results. General background information was provided about the data and tasks. This was particularly important for the Phylogenetic data.

4.3.3 The 2003 Submissions and Judges

Teams had about five months to prepare. We had several judges in standby mode in case we received a large number of submissions. We received eight entries, a small number, but satisfactory for a first contest. The judging was completed by the two chairs and two additional judges, all knowledgeable in information visualization and human-computer interaction, with one of the judges also knowledgeable in biology, useful because of that specific dataset.

4.3.4 The 2003 Results

The first main finding was that the tasks and datasets were too complex for such a contest. Each tool addressed only a subset of the tasks and only for a subset of the datasets. The phylogeny chosen required domain expertise hence was "real", and even though it consisted of a small binary tree, it was not used, probably because the tasks were complex and required biological knowledge (e.g., perhaps working with a biologist).

The second main finding was that it was difficult to compare systems even with specific datasets and tasks. We had hoped to focus the attention of submitters on tasks and results (insights), but the majority of the materials received focused on descriptions of system features. Little information was provided on how users could accomplish the tasks and what the results meant, making it very difficult for the judges to compare results. The systems presented were extremely diverse, each using different approaches to visualize the data.

Selecting the three first-place entries was straightforward. Only the first place submissions demonstrated the benefits of their tools by reporting on the insights gathered, and both the descriptions of the tool and processes were clear making it easy to understand how the tool had been used (see Figure 3).

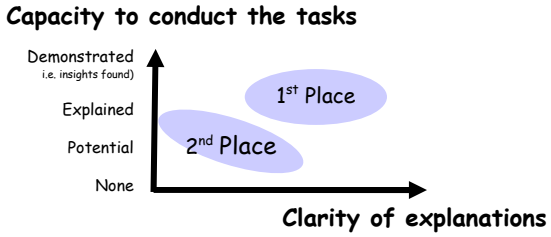

Figure 3: General comparison of 1st and 2nd place entries

TreeJuxtaposer [Mun03a] (Figure 4) submitted the most convincing description of how the tasks could be conducted and the results interpreted (see also [Mun03b]). Zoomology [Hon03] (Figure 5) demonstrated how a custom design for a specific single dataset could lead to a useful tool that addressed many of the tasks satisfactorily. InfoZoom [Spe00] (Figure 6) was the most surprising entry. This tool had been designed for manipulating tables and not trees. However the authors impressed the judges by showing that they could perform most of the tasks, find errors and provide insights in the data. The three second-place entries showed promise but provided less information to the judges on how the tasks were conducted and the meaning of the results. EVAT [Aub03] (Figure 7) demonstrated that powerful analytical tools complementing the visualization could assist users in accomplishing their tasks. TaxoNote [Mor03] (Figure 8) demonstrated that labeling is an important issue making textual displays attractive. The Indiana University submission [She03] (Figure 9) illustrated the benefits of toolkits (e.g. [Bor06; Fek04] by quickly preparing an entry combining several tools, each accomplishing different tasks.

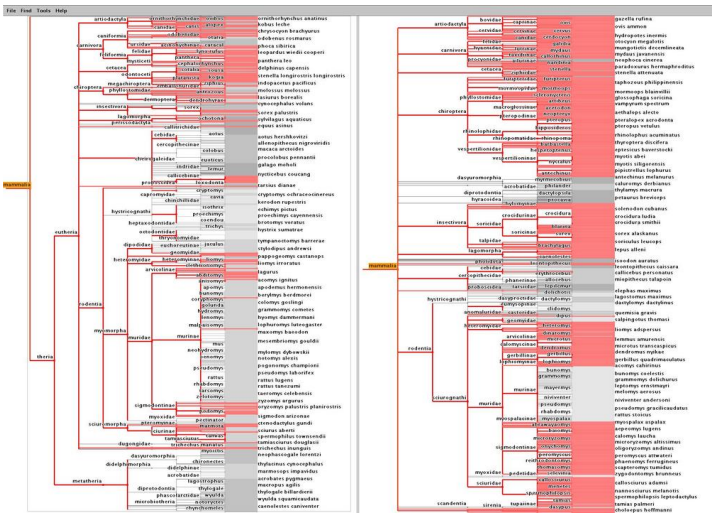

Figure 4: TreeJuxtaposer [Mun03a]

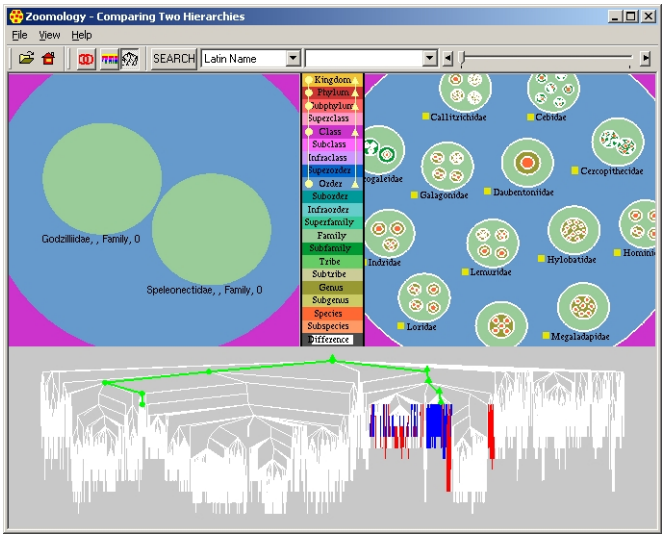

Figure 5: Zoomology [Hon03]

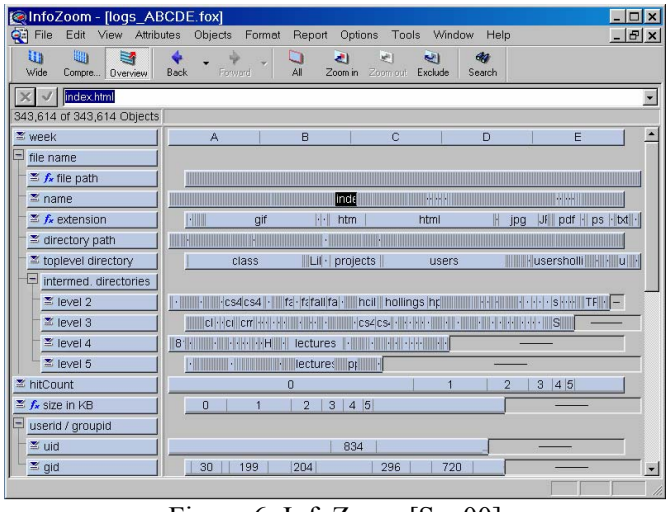

Figure 6: InfoZoom [Spe00]

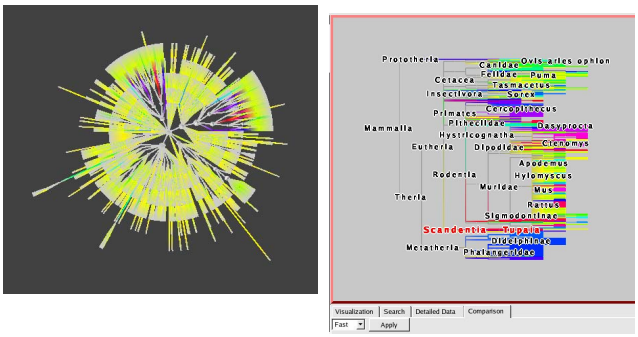

Figure 7: EVAT [Aub03]

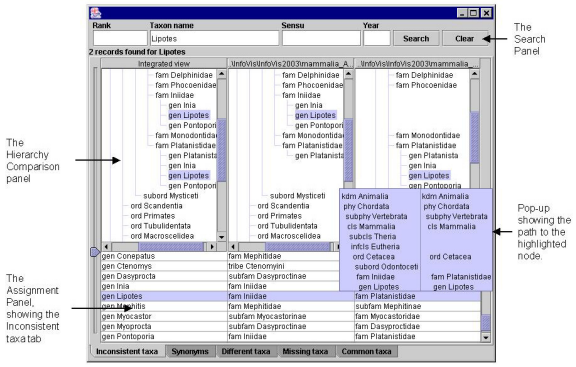

Figure 8: TaxoNote [Mor03]

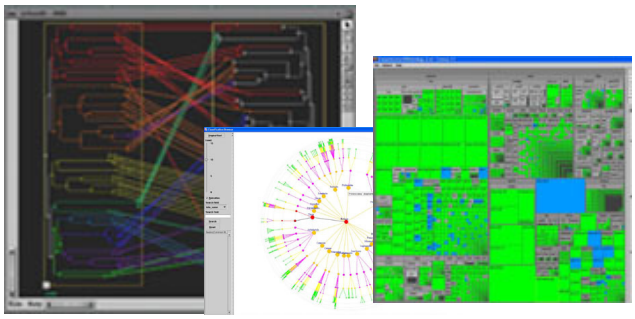

Figure 9: A combination of tools - Indiana University [She03]

All entries were given a chance to revise their materials after the contest. Participants provided a structured form with screenshots and detailed explanations for each task.

4.4 The InfoVis 2004 Contest

The second competition coincided with the 10 year anniversary of the InfoVis Symposium. As the visualization of the history of a research field is a problem interesting in itself, it was natural for it to become the core part of the contest. The key advantage of the topic was that it was familiar to all participants. The disadvantage was that the selected corpora were not readily available in a usable form.

4.4.1 The 2004 data

The set of all publications on a topic is too large a universal set of discourse for a competition. We first

argued about which conferences or journals to include and resolved to limit the dataset to all the IEEE InfoVis Symposium papers and all of the articles used as reference in those papers. Metadata is rich for IEEE and ACM publications and unique keys available.

Producing a clean file (metadata for the collection of documents) was a much bigger challenge than we had imagined. We first made an assumption that both the articles and the most important authors in information visualization would be referred by most of the articles published within the InfoVis symposium. Our look at the references initiating from articles published within InfoVis seemed to us at the same time focused on the field and complete. It would be unlikely that an important publication in information visualization would seldom be referenced by other articles.

This was partially correct but text metadata still yielded numerous ambiguities. IEEE manages the InfoVis articles which are less curated than those of the ACM. Much text metadata was non-unique (e.g., many-to-one names such as Smith, Smyth, Smithe, …). Reference titles were too inaccurate and in many cases erroneous as text handled by the ACM Digital Library uses strings and numerical computations with such string are still imprecise (long string comparisons). Much curation on our end was necessary as references were noisy, sometimes missing, and even sometimes pointing to non-existing URLs.

We thus embarked on cleaning the data. This was a complicated process, with multiple passes, and manually intensive, even with automatic reference extraction as we found no reasonable automatic system to suitably resolve the problems. We manually extracted the articles from eight years of pdf files from the symposia available in the digital library. We then semi-automatically retrieved the articles referenced in those papers again from the digital library. We extracted those which existed when found and manually cleaned and unified the publications not included in ACM library.

The result was a file containing 614 descriptions of articles published between 1974 and 2004 by 1,036 authors, referencing 8,502 publications. It took well over 1,000 hours for us to construct that file, with over 30 people involved.

4.4.2 The 2004 tasks

We proposed 4 high level tasks with a great deal of flexibility for a variety of solutions:

1. Create a static visualization showing an overview of the 10 years of InfoVis
2. Characterize the research areas and their evolution
3. Explore the people in InfoVis: where does a particular author/researcher fit within the research areas defined in task 2?
4. Explore the people in InfoVis: what if any, are the relationships between two or more researchers?

We suggested particular names for task 3 to facilitate comparisons between submissions, and participants used them, along with other names.

4.4.3 The 2004 submissions and judges

The participants were required to submit:

- a two page summary,
- a video illustrating the interactive techniques used,
- a structured web form providing details as to how the tasks were accomplished and what discovery or insights were identified.

There were 18 submissions from 6 countries (USA, Canada, Germany, France, Australia, and Netherlands) with 7 academic participants. The judging was done by the three contest chairs (the authors) with the help of one outside reviewer. Twelve teams were selected to have their results published in the InfoVis repository. Four received a first place prize and gave a short talk at the conference. Six teams received a second place prize and presented a poster.

4.4.4 The 2004 results

Quality improved dramatically between 2003 and 2004. The good news was that most teams had provided a lot more insights than we had seen in the first contest. Still, some teams had tools that seem promising "on paper" but reported very few insights (in consequence they did not do very well in the contest.) On the other hand some teams presented tools that seemed of doubtful utility to the reviewers at first but were able to report useful insights, therefore faring better than we had expected in the results. Of course the strongest teams performed well with all requests: had promising visualizations, many insights reported, and convincing explanations of how the insights were obtained using the tools.

None of the twelve selected teams answered all the questions. A few of the participants had extensive experience with text analysis and that was visible in their results. Other had background knowledge of the InfoVis community and could provide better hypotheses about what they were seeing. One tool was developed entirely from scratch for the contest but most teams showed interesting new uses of existing techniques. Node-link

diagrams were a very commonly used representation for many of the tasks, with some notable exceptions.

This second contest had a single dataset and simpler tasks so we anticipated reviewing and comparing results would be much easier. Not so. Again, we had hoped that we would be able to evaluate the quality of answers computationally but this was not possible. The problem again was the fuzziness of the answers and the lack of "ground truth" or even consensus on what the best answer might look like. Teams' answers took various forms from a collection of articles or names to a new algorithm to a new visualization, all of whose correctness was not computable. Only human evaluation was appropriate to judge the validity of the answers. In information retrieval, TREC for example does use human judging and pools results from all participants to determine the relevance of documents (i.e. the answers) from which metrics can be computed for a team's set of results. Short of spending time with the team throughout the discovery process (an extremely interactive and personal activity) we could only base our judgment on the materials provided (the video and forms).

There were three first place entries and one student first place:

- The entry from Indiana University [Wei04] had many insights gathered from a variety of mostly low tech displays such as barcharts, displaying the output of separate analysis tools. The best display showed a simplified view of the co-authorship network clearly highlighting the main players in the field of InfoVis (as represented by the dataset) (Figure 10). It has since then been used by many as an overview of the community. The entry from the Pacific Northwest National Laboratory [Won04] nicely reflected the extensive experience of the team and the rich set of analysis features provided by their tools (Figure 11).

- The entry from Microsoft and the University of Maryland [Lee04] was interesting because it took a completely different approach to the problem, departing from node-link diagrams, and using multiple tightly coupled ordered lists to represent the interconnections between authors, papers and references (Figure 12).

- The student first prize went to a team from the University of Sydney [Ahm04]. It surprised us with its whimsical 3D animation and provided sufficient insights to convince us that the technique had merit (Figure 13).

Second place prizes (see Figure 14 to 20) went to the Université de Bordeaux I with the University of British Columbia [Del04], the Technische Universiteit Eindhoven [Ham04], Georgia Institute of Technology [Hsu04] and [Tym04], the University of Konstanz [Kei04], the two teams from Drexel University [Lin04] and [Chen04], and the University of California, Davis [Teo04]. Each has some interesting technical component and some interesting insights.

We were satisfied that teams reported useful insights but we were still surprised by how few were reported, and that even fewer were really surprising insights. Insights about the whole structure were rare and only came from teams who had experience looking at other domains (e.g., the fact that InfoVis is a small world, tightly connected, was mentioned only by 2 teams.) One team noticed that the most referenced papers were published at CHI, not at InfoVis. Only three teams noticed the existence of references to future papers, a problem resulting from automatically processing references and confusing multiple versions with similar titles such as a video and a paper. Only one insight highlighted something surprisingly missing, namely that there were no papers in the dataset from several of the other competing InfoVis conferences, despite the fact that they had been held for several years.

Teams interpreted the tasks and used the data in surprisingly very different ways. A task such as "describe the relationships between authors" was interpreted in at least the following nine different ways as report on co-authorship; or co-citation; or people working on similar topics; or having a similar number of co-authors; or being a part of big groups or teams; or having a similar number of publications; or a similar number of references; or working in the same institution; or working in the same clique or empire.

Teams also used the data differently. Some created displays showing either only the IEEE Infovis symposium papers or all papers including the references. Some combined both authors and topics and some used separate displays. In one case we suspected that a team used only the papers' first authors but were unable to determine that precisely. One team only used references from InfoVis papers references but not references to papers from other venues.

The presented data was generally pruned dramatically to work with the tools or to create more useful or possibly appealing displays. Few attempted to show complete views. Some teams had a "celebrity" approach ignoring everything but the star papers or authors based on some unique criteria (e.g., numbers of citations). Some clustered first then pruned later with no clear explanation of what had been pruned.

Reviewing the displays seemed easy at first, but it quickly became impossible to remember what data we had just been looking at, let alone compare different results even when it would have been possible. Many displays had no or very poor captions and none had any summary of the process that generated the display. Each team probably had a clear model of the scope of the data and how it was filtered, aggregated and interpreted, but the displays did not reflect that.

Few teams even attempted to answer the first question, to create a static visualization showing an overview of the 10 years of InfoVis. Teams merely reused one of their screen shots from other tasks so we felt only one aspect of the data had been portrayed and not the entire 10 years of InfoVis. Teams reported very different topics and different numbers of topics (from 5 to 12) and some created topics on the fly, refining the topics iteratively. Sometimes a seemingly narrow topic would take a prominent place: "parallel coordinates" was a major topic in one case while in another system "taxonomy" was a major topic. Reviewing all the submissions gave us an impression of randomness in the choice or labeling of the topics. One of the student teams used their professor's notes to extract topics. It was innovative but again, affected our ability to make comparisons. Most visualizations limited the total number of topics which limited the insights to be related to those topics. But topic extractions were not the focus of the contest so we did not judge the quality of the topics. Nevertheless this made it more difficult to compare insights. Some tools (e. g., In-spire [Won04] and an entry from Rutgers University [Lin04]) allowed users to refine the topics quickly and iteratively by entering a seed term or removing words. Some teams chose to remove common words (e.g., Information Visualization) while others kept them as labels for major topics which was not very useful. Sadly, evaluation only appeared once as a topic, and only after iterative refinement.

Overall, labeling remained a very big problem. Very rarely could we actually guess paper titles when looking at a display. Better dynamic layout techniques for labels were clearly needed. Labels for papers usually consisted of the first few words or even just the first author making it difficult to remember if we were looking at author relationships, or papers, or even topic relationships, e.g. a large node labeled "Johnson" could represent the often-referenced Treemap paper. Part of the problem was simply that teams most often used exploratory tools for discovery but use the same displays for presentation.

Some tools used a single window [Teo04], but most used multiple windows, showing either variants [Ham04] or very different displays for different tasks [Chen04], [Kei04]. The PaperLens submission [Lee04] illustrated the importance of coordinating views. Only two teams dealt with missing data and uncertainty, others ignored the problem entirely. Visual metaphors seemed to have had an effect on the words teams used to describe their findings, e.g. one team [Ahm04] talked about empires when looking at towers in 3D, while others talked about cliques while looking at clusters on node link diagrams. Unfortunately, we also saw examples of "nice pictures" that didn't seem to lead to any insight.

The 2004 contest session at the workshop was very well attended and we received extremely positive feedback. Attendees reported an appreciation for the wide diversity of solutions and contrasting different techniques. We conjecture that the topic we had selected also made the contest more appropriate to the audience as well as more accessible.

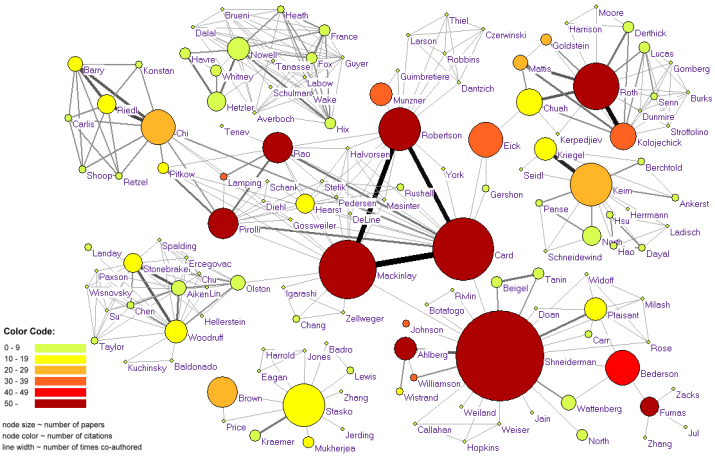

Figure 10: Node-Link diagram, from Indiana University [Wei04]

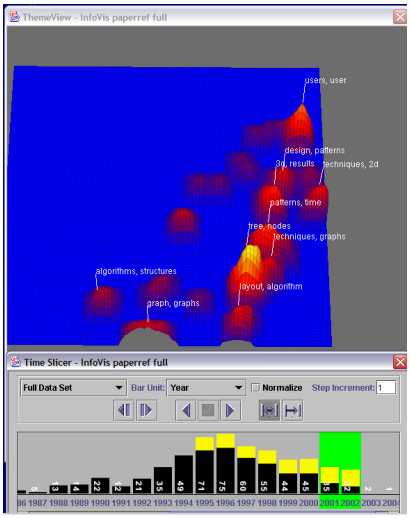

Figure 11 In-Spire clusters from the Pacific Northwest National Laboratory [Won04]

Figure 12 PaperLens distributions from Microsoft and the University of Maryland [Lee04]

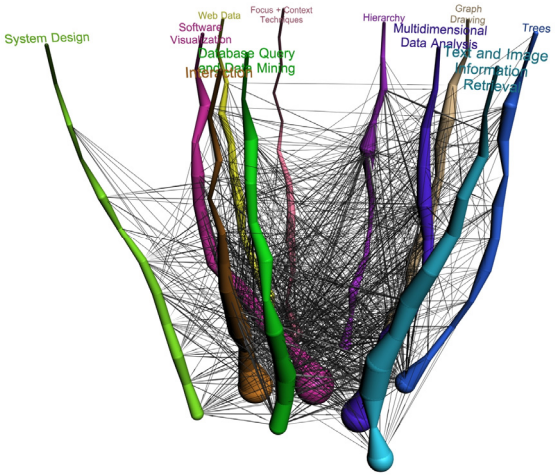

Figure 13: Wilmascope topic flows from the University of Sydney [Ahm04]

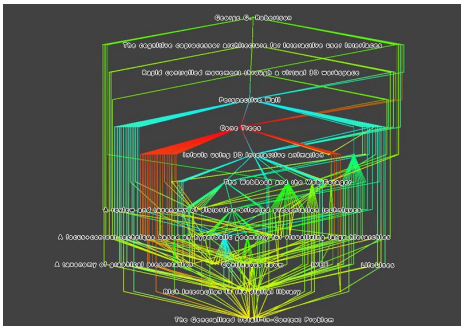

Figure 14: Document graphs from the U. de Bordeaux I and the University of British Columbia [Del04]

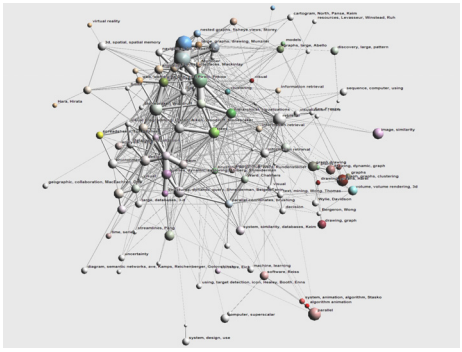

Figure 15: Node-Link diagram from the Technische Universiteit Eindhoven [Ham04]

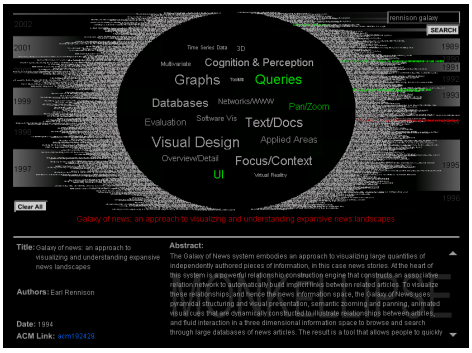

Figure 16: Topical overview and focus from Georgia Institute of Technology [Hsu04]

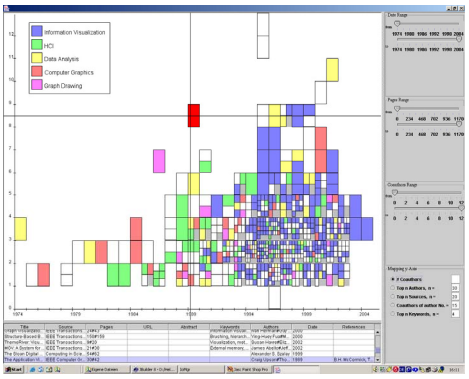

Figure 17: Document timeline and classes from the University of Konstanz [Kei04]

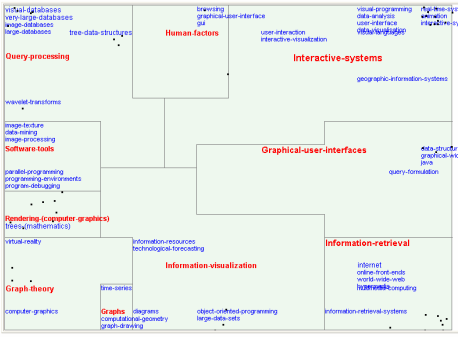

Figure 18: Topic classification from Drexel University [Lin04]

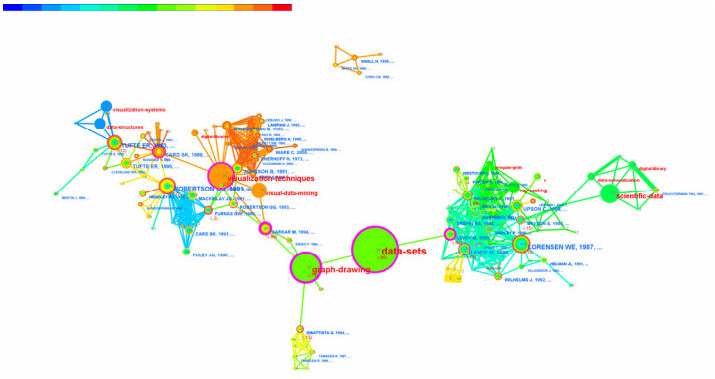

Figure 19: Author link diagram from Drexel University [Che04]

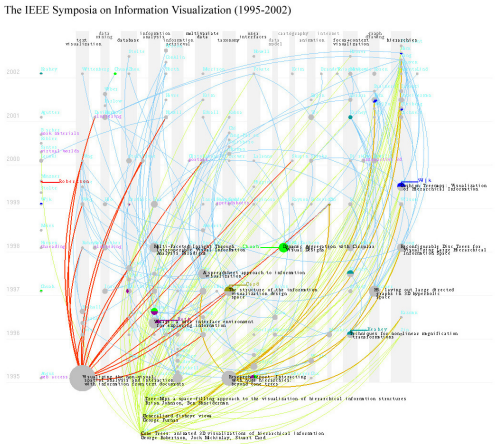

Figure 20: Topic and author timeline from the University of California, Davis [Teo04].

4.5 The InfoVis 2005 contest

In the third competition [Inf05] the chairs aimed for the evaluation of more complete visualization systems and a different type of data. The data set was larger and the questions more targeted. The goal was to identify how well visualization or visual analytics systems or even specific tools could perform with a large but easily understood data set. The chairs missed a key point in that the problem was probably better phrased as a Geographic or Geospatial Information (GIS) challenge rather than simply an information visualization one. The chairs also released the data set only for the competition. The owner of the data set did not permit an open release, something that the chairs tried to avoid and hopefully will avoid in the future.

4.5.1 The 2005 data

This large, information rich, and real data set consisted of information on 87,659 technology companies in the US, and included the year founded, zip code, yearly sales, yearly employment information, along with industry and product information using the North American Industry Classification System [Nai01]. This was a large data set with geographic interpretation, one which pushed the limits of many systems. The data was cleaned by graduate students at the University of Massachusetts at Lowell. Missing data was eliminated as best as possible.

4.5.2 The 2005 Tasks

The three questions related to the characterization of correlations or other patterns amongst variables in the data were

1. Characterize correlations or other patterns among two or more variables in the data.
2. Characterize clusters of products, industries, sales, regions, and/or companies.
3. Characterize unusual products, sales, regions, or companies.

One additional question was more general and open-ended

4. Characterize any other trend, pattern, or structure that may be of interest.

The chairs felt that these precise questions would make evaluation simpler. And again this was not correct as all questions were too open-ended in their interpretation and comparing the discovery of different correlations was difficult.

4.5.3 The 2005 Submissions and judges

Participants were required to submit materials using the same format as in 2004. There were only 10 participants. This was a surprise but the short time from available data to submission deadline and the size of the data were probably the most important factors. We had no submissions from student teams possibly because we released the first version of the data set at the end of February during which most university information visualization classes already are well under way. The judging was done by the four contest chairs using some numerical ratings but again the selection process was fairly straightforward.

4.5.4 The 2005 Results

The chairs managed the review process and evaluated the entries in a similar manner as the previous year, but used specific measures for insight, presentation, interaction, creativity, flexibility, and novelty. There were two first and two second place awards. Teams led by the Iowa State University [Hof05] and Penn State

University [Che05] took first place having answered all questions, while the Universität Karlsruhe [Hos05] and Augsburg University [Zei05] provided strong answers and received second place prizes.

The first place winners took two different approaches. The team from Iowa State University provided a classic information visualization approach and highlighted the important role of the analyst [Hei05]. Their system did not have high end visualizations but their analyses led to answers for all questions (Figure 21). Their focus on "boom and bust" was right on target and very elegant. The Penn State University and University of South Carolina team looked at the problem from the point of view of a GIS time-based system and used many of their map visualization tools to explore cross-state patterns (Figure 22) [Jin05]. Their system was the most novel with superb interaction. This of course highlighted the duality nature of the data set. One could look at it strictly from the information visualization point of view or from the GIS point of view and thereby harness a great deal of knowledge and techniques from the mapping and geospatial fields.

The second place winners had strong answers. The Company Positioning System from the Universität Karlsruhe was visually stimulating, and had high scores on interaction and novelty (Figure 23) [Hos05] while the team from Augsburg University used interactive statistical graphics to derive and present their answers (Figure 24) [Zei05]. Their interactive solutions were insightful.

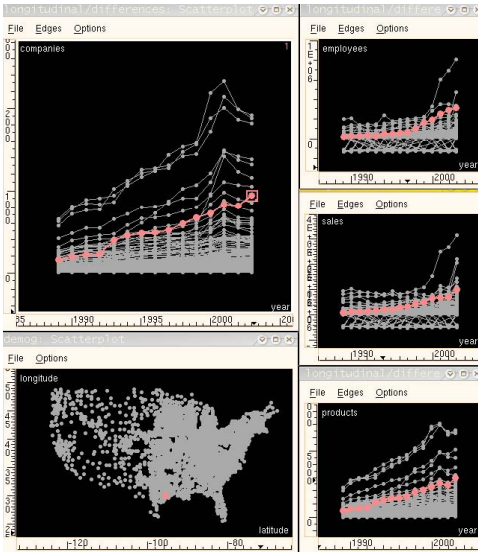

Figure 21: Iowa State University [Hof05]

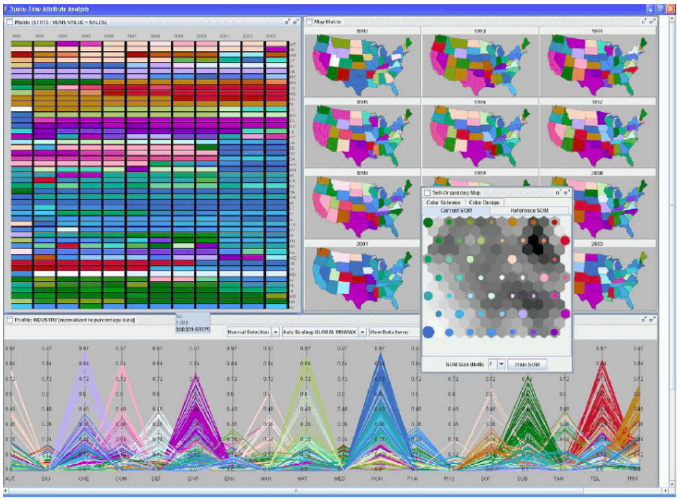

Figure 22: Penn State University [Che05]

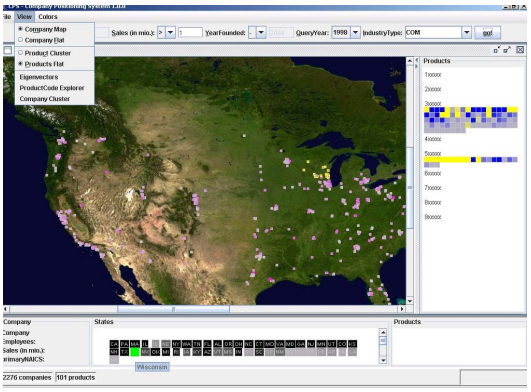

Figure 23: Company Positioning System by the Universität Karlsruhe [Hos05]

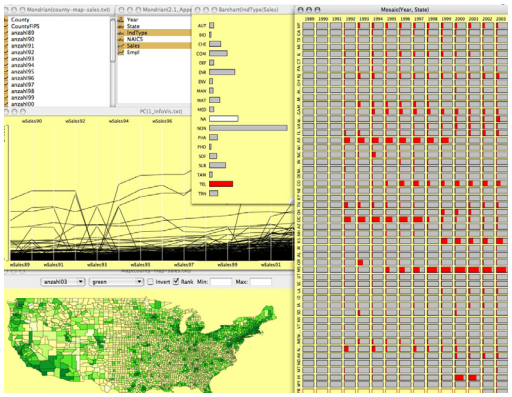

Figure 24: Augsburg University [Zei05]

All in all the four winners covered a broad spectrum of techniques for information visualization solutions. In all cases we found that statistical and computational methods played a key role. The data was just too large for simple human consumption thereby putting visualization in a collaborating role with analysis. Much processing of the data took place and all contestants used coordinated views to answer the questions.

4.6 Lessons Learned from the Three Contests

4.6.1 About the evaluation process

The contests illustrated the difficulty researchers have at giving evidence that their tools can effectively perform the tasks. Demonstrating the power of a tool can be difficult. Researchers are trained to describe their tools' novel features more than illustrating them with convincing examples using real data. In 2003 participants barely reported any insight at all. Everyone was focused on the description of their tools. By 2004 more participants (not all) were able to provide insights. In 2005 insights were more common. Providing good examples of insight reports seemed to help the participants.

Half of the participants were students who built their own tools. These tools were not as polished as industrial products or well developed research systems. Providing benchmarks that fit student project's sizes seems important to the success of future contests. In 2003, we provided large data sets with some meaningful subsets. In 2004 the data set was not very large. However, in 2005, the dataset was large and there was no subset provided so the number of participating students dropped.

The evaluation process is time consuming and looked at by some of the chairs as a daunting task. Ideally one could compute metrics and add these up assuming independence to get a summary score, but since the questions are open-ended and there is no known "ground truth" human evaluation is still required. Even though there were a limited a number of specific tasks, these can be interpreted differently and of course participants did interpret them in many ways. We always felt as if we had to compare non-comparable processes and results.

Evaluating the results remained a subjective activity. After seeing the submissions, the 2003 and 2004 contest reviewers decided to classify the teams into three categories based on increasing insight: no evidence of insight gathered using the tool, some insight, and much insight, i.e. worthy of a first place. On the other hand the 2005 contest used more than 6 categories. This helped with discussions but required a great deal more time reviewing the materials, and it is not clear that it resulted in better selections. This is an issue, as we do not have a good way to evaluate insight without ground truth. The recently completed visual analytics contest associated with the IEEE VAST 2006 symposium [VAS06; Gri06] provided more evidence for the benefits of having ground truth. It introduced a promising approach for the development of automated tools and the collection of metrics.

Another challenge in judging is the required repetitive, often mostly visual evaluations of similar entries. It can be very hard to remember "who did what" or "who had this insight". An insight implies novelty of the finding, so reviewers might more positively weigh a reported insight the first time they encounter it and undervalue it later on when reported by another team. Having a shared environment to manage the tracking and comparison of the reported insights would be every helpful.

Videos were extremely important. Without them it would have been impossible to understand how most tools worked and what process was used to answer the questions. With videos, interactions become understandable. Verbal comments on the videos were indispensable in explaining what the participants were doing. This is quite different than simply reviewing a paper. On the other hand video requires the reviewers to remember key points as it is quite difficult to scan a video quickly to refresh one's memory. Dealing with videos was also very time consuming. For the first two contests we were flexible about the format of the videos submitted but this created problems such as finding format converters or hunting for missing codecs. For the 2005 contest we required a single format which simplified the process. Requesting multiple short videos may also prove helpful.

4.6.2 Data preparation

With the InfoVis 2003 contest we attempted to provide real data and tasks while trying to narrow the problem to one data type (trees) and three representative tree types. The contest taught us that the problem was still too large for a contest and that the vague nature of the tasks made it impossible to compare answers effectively. In contrast the 2004 contest had only one dataset, much fewer tasks and a more structured reporting format. Nevertheless, the open-ended nature of realistic tasks and the diversity of interpretations and approaches still made judging the submitted entries a challenge.

We felt that the time to generate a reasonably clean data set was too large, around 1000 staff-hours each year with a similar large figure for the VAST 2006 and 2007 contest data sets. This is a serious issue for the development of benchmarks. Domain experts should be solicited for cleanup and experimentation on various task solutions should be attempted before the data is released. We hope that industry groups or government agencies wishing to see more research conducted on specific data of interest to them will take on the burden of developing the benchmarks datasets or support groups to do so.

4.6.3 Motivation of participants

Participating in the contest takes time and motivation. Most participants reported working very hard to prepare their submission. Many acknowledged that it pushed them to improve and test their tools. Students were encouraged to work on the contest as class projects sometimes continued beyond the semester. Some wanted to test their PhD research. Small companies reported appreciating the exposure.

In all three contests we gave small prizes from various sponsors to the first and second place teams. All appreciated them especially the students who received gaming systems. We also presented all winners with certificates. Participants appreciated being able to mention the award and the resulting publication (even if a minor one) in their resumes. On the other hand, in our polling (see below) some tenure-track faculty reported being interested in the contest but preferred focusing on writing full papers instead. Providing better incentives should help attract more participants.

We felt that we should have chosen the data set and plan for an earlier release date. Given that we ran into errors and noise in almost all the data sets, having more time might help clean the data and prepare better tasks. Pushing the deadline further into the late summer would allow summer interns to work on the contest, but would reduces the reviewing period dramatically, which doesn't seem feasible for a conference in October.

Many people downloaded the dataset without submitting results and we collected names and emails (for 2004 and 2005). The chairs performed an informal survey of those that had downloaded the 2005 contest to see why there were so few participants. The participants stated that there were no problems with the data set or questions, that the data set was a great data set to show system and tool capabilities, and that all had enjoyed the process and would do it again. Most expressed that they wanted a better organized website, automated email on data or news updates, and would have preferred the data in a database format. Some expressed strong interest in splitting entries into commercial and academic categories. Four of the non-participants stated that the requirement to attend the conference hindered their participation and most expressed that they were too busy in their company to tackle such a project. Several expressed a desire for mini-questions such as "find a more elegant way to look at …"

There was one recurring theme which all participants and non-participants expressed and that was the need for more time. That was the reason the 2006 contest data was made available at the 2005 conference, but again the participation did not increase substantially so other factors may be more important, such as incentives, or personal interest in the data and problem proposed.

4.6.4 Running a successful event at the conference

For all three contests we had a full session at the conference where we summarized the results and had some authors present their submissions. In 2003 only the first place authors presented and we summarized the second place submissions. Many commented that it would be better to have shorter presentations but allow more presenters to speak. The following year we arranged for all first and second place winners to present with the second place ones having only 2 minutes. This format was very well received. We specified tasks that presenters should focus on so the attendees could better compare the different entries, at the cost of not seeing every feature of the tools. We found that handing out the awards rapidly and keeping photos to a minimum (a group picture at the end of the session) was preferable. This left more time for the presentations and still gave a festive atmosphere to the event. All winners were also given a chance to have a poster displayed during the normal poster session.

4.7 Impact and repository

A contest is only a first step. The revised materials provided by the authors and the datasets have to be made available after the event to be actively used. We are keeping the contest pages active and have made the submissions available in the InfoVis repository hosted at the University of Maryland [Bmr06]. On that website we encourage researchers and application developers to continue using the datasets and tasks. We hope that others will add the results of their analysis to the repository thereby enriching it and providing a more comprehensive review of visualization techniques available. Appropriate advertising and promotion of the repository should help. More importantly we believe that an infrastructure is needed to support and facilitate the use of those datasets (e.g. by providing multiple views and versions for different uses) by monitoring the usage of the repository (downloads, visits) and gathering information regarding its impact (e.g. follow up to collect researchers' reports on their use of the datasets, or collections of the citation). In 2003 we did not monitor any activity; in 2004 we know that the dataset was downloaded over 350 times by the end of 2006, and we found 16 papers that mentioned using the 2004 contest datasets (e.g. [Bor06b, Fai06, Min06]), but a more thorough follow-up would be needed to evaluate the impact of the contest. The 2005 contest is now making its appearance in publications ([Unw06], [Che06a], [Che06b], [Guo06]). Running an information visualization contest is fairly taxing and volunteers who

run the contest tend to run out of steam by the end of the contest, making reporting of lessons learned and follow-up activities difficult. We believe that modest support for a long term coordinated evaluation program and infrastructure would greatly increase the impact of the information visualization contest and the benchmark repository.

5 Recommendations

We now propose a set of practical recommendations for tools researchers, contest organizers, and visualization users.

For potential contest participants

- Find partners if you cannot complete the task alone
- Suggest next years' topic if this year doesn't fit your research or directions
- Encourage your sponsors to prepare data sets for contests
- Ask questions (you rarely ever do)
- Participate even after the contest by adding entries to the repository

For researchers

- Exercise your tools with the benchmarks datasets and tasks
- Participate in contests (those who participate report many benefits)
- Leverage existing contests to test your new metrics or novel evaluation methods

For organizers of contests

- Provide incentives for contestants that go beyond gifts (e.g. negotiate full papers in journals or plan early for the inclusion of the summaries in a digital library). Announce those incentives in the call. This may be the most driving force for academics, students and faculty alike.
- Facilitate student participation (choose the schedule carefully, advertise in classes, include smaller problems, plan for travel grants or student volunteer positions)
- Provide templates and examples for participants to report their findings in a structured fashion
- Set up a registration process to monitor downloads and plan to monitor and encourage usage after the contest. Plan to write reports in a timely fashion.
- Use datasets with established ground truth
- Balance the diversification of problems chosen with some continuity from year to year to allow participation to build up
- Establish connections with funding agencies to plan long term evaluation activities

For potential visualization users

- Prepare sanitized (if necessary) and interesting benchmark datasets and tasks (or support researchers to prepare them)
- Offer your help to generate ground truth for benchmark datasets
- Participate in contests using the off-the-shelf technology you use today, to provide reference points
- Encourage and support contest and evaluation activities in general

6 Conclusions

Although a contest is an artificial testing situation the information visualization contests encouraged participants to thoroughly exercise their systems over a long period of time, mimicking a fairly realistic analysis process. Participants were asked to report on the insights gained from exploring the data. The impact of contests is most obvious for those who participate and those who can compare their results at the conference but the datasets, tasks and results remain available after the contests thereby extending their impact. Contests can be used by developers to exercise their tools and identify missing or weak features, and to compare their tools' effectiveness with those of others. Evaluators can use contests to evaluate, explore and enrich their testing procedures with complex tasks. Developers and evaluators then will have baseline results with which to compare their own results. We hope that these data can also be used in controlled experiments, and that the other more specific lists of tasks used in those experiments will be added to the repository for further reuse.

Benchmarks are difficult to create, promote and use. Our belief is that we are developing solid and evolving benchmarks and are beginning to understand how to better evaluate submissions. Good benchmarks must be real (witness the success of TREC and CAMDA, and the excitement generated by the VAST contest) to both draw the audiences and participants and to strongly push the technology curve. Good benchmark tasks must be open-ended to provide for the flexibility in solutions. We know that this makes the results more difficult to measure analytically but this is realistic. We need to accept that more human evaluation will be required in the future and evolve a collection of volunteer judges. These contests continue to demonstrate the challenges of benchmark design and especially of system and tool evaluation.

By making the results of analyses available to the community in a benchmark repository we provide a set of comparison baselines for developers. Even though teams did interpret our tasks in many different ways, making comparison difficult, we feel strongly comparison with the same data set and tasks are useful and important. Plus the different number of interpretations could be reduced by more explanation.

The integration of analysis is now more necessary as data sets are more complex, large, and coming from diverse sources. The identification of anomalous patterns of data from phone calls, from bank transactions, and from news articles requires new techniques and strong analytical tools. We believe that such data sets and competitions will continue to encourage the community to work on difficult problems while building baselines of comparable tasks and datasets.

Acknowledgements

For all the contests, we thank the organizers of the IEEE Visualization InfoVis Symposium, in particular John Dill, Tamara Munzner and Stephen Spencer, members of the InfoVis community for their intellectual stimulations, and the participants without whom there would be no contest. We would also like to thank the numerous students for their help in extracting the metadata, cleansing the data set, and producing a richly usable data set: Caroline Appert (Université Paris-Sud, France) and Urska Cvek, Alexander Gee, Howie Goodell, Vivek Gupta, Christine Lawrence, Hongli Li, Mary Beth Smrtic, Min Yu and Jianping Zhou (University of Massachusetts at Lowell). After the first release of the datasets many others offered their help, including Jeff Klingner from Stanford, Kevin Stamper, Tzu-Wei Hsu, Dave McColgin, Chris Plaue, Jason Day, Bob Amar, Justin Godfrey and Lee Inman Farabaugh, from Georgia Tech, Niklas Elmqvist from Chalmers, Sweden, Jung-Rung Han, Chia-Ning Chiang, and Maylis Delest from the Université de Bordeaux. We want to thank Cynthia Parr, a biologist at the University of Maryland and Elie Dassa from the Institut Pasteur in Paris for helping us create the 2003 datasets. We also thank Anita Komlodi from UMBC and Cyndy Parr for participating in the review process. We thank ACM and IEEE and in particular Mark Mandelbaum and Bernard Rous for helping make the 2004 data available and working with us to prepare the dataset, Shabnam Tafreshi for help with the website, and finally but not least Paolo Buono from the University of Bari, Italy, for participating in the review process. We thank Michael Best (University of Massachusetts at Lowell) for working with us on releasing the technology company data for the 2005 contest. We thank the sponsors who provided prizes, The Hive Group, ILOG and Stephen North in 2003, Kitware, Nvidia and Microsoft research in 2004. We also thank Sharon Laskowski for working with Catherine Plaisant on the evaluation section of the NVAC research agenda [Tho05] which helped refined some of the sections of this paper and lead to Figure 2.

References

[Ahm04] Adel Ahmed, Tim Dwyer, Colin Murray, Le Song, Ying Xin Wu, WilmaScope, *Poster Compendium of IEEE Information Visualization* (2004)

[Aub03] Auber, D., Delest, M., Domenger, J-P., Ferraro, P., Strandh, R., EVAT - Environment for Visualization and Analysis of Trees, in *Poster Compendium of IEEE Information Visualization* (2003)

[Bel06] BELIV'06, BEyond time and errors: novel evaluation methods for Information Visualization, a workshop of the AVI 2006 International Working Conference. http://www.dis.uniroma1.it/~beliv06/

[Bmr06] Information Visualization Benchmark Repository www.cs.umd.edu/hcil/InfovisRepository

[Bor06] InfoVis CyberInfrastructure — http://iv.slis.indiana.edu

[Bor06b] Börner, K., Dall'Asta, L., Ke, W. and Vespignani, A., Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams. *Complexity*, Special issue on Understanding Complex Systems 10(4) (2005) 58-67

[CAM06] Critical Assessment of Microarray Data Analysis (CAMDA) conference, http://www.camda.duke.edu/camda06

[CES06] Council of European Social Science Data Archives (CESSDA) – http://www.nsd.uib.no/cessda

[Che00] Chen, C., Czerwinski, M. (Eds.) Introduction to the Special Issue on Empirical evaluation of information visualizations, *International Journal of Human-Computer Studies,* 53, 5, (2000), 631-635.

[Che04] Chen, C., Citation and Co-Citation Perspective, *Poster Compendium of IEEE Information Visualization* (2004)

[Che05] Jin Chen, Diansheng Guo, Alan M. MacEachren, Space-Time-Attribute Analysis and Visualization of US Company Data, *Poster Compendium of IEEE Information Visualization* (2005)

[Che06a] Chen, J., MacEachren, A. M., & Guo, D. 2006, Visual Inquiry Toolkit An Integrated Approach for Exploring and Interpreting Space-Time, Multivariate Patterns. AutoCarto 2006, Vancouver, WA, June 26-28, 2006 (CD only).

[Che06b] Chen, J., MacEachren, A. M., & Guo, D. in press, Supporting the Process of Exploring and Interpreting Space-Time, Multivariate Patterns: The Visual Inquiry Toolkit. Cartography and Geographic Information Science.

[Chi93] Chinchor, N., Hirschman, L., Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3), *Computational Linguistics* 19, 3 (1993) 409 - 449

[Cow05] Cowley, P., Nowell, L., Scholtz, J., Glassbox: an instrumented infrastructure for supporting human-interaction with information, *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (HICSS'05) , pp. 296.3 (2005)

[Del04] Maylis Delest, Tamara Munzner, David Auber, Jean-Philippe Domenger, Tulip, *Poster Compendium of IEEE Information Visualization* (2004) [Fek03] Fekete, J-D and Plaisant, C., InfoVis 2003 Contest, www.cs.umd.edu/hcil/iv03contest (2003)

[Fai06] Faisal, S., Cairns, P., and Blandford, A., Subjective Information Visualisations. In *Proc. Workshop on Combining Visualisation and Interaction to Facilitate Scientific Exploration and Discovery*, held in conjunction with the *British HCI 2006* conference, London, UK (2006)

[Fek04] Fekete, J.-D., The InfoVis Toolkit. in *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*, (2004), IEEE Computer Society, 167-174.

[Geh04] Gehre, J., Ginsparg, P., Kleinburg, J., Overview of the 2003 KDD cup, *SIGKDD Explorations*, 5,2 (2004) 149-151.

[Gon03] Gonzales, V., Kobsa, A., Benefits of information visualization for administrative data analysts, *Proceedings of the Seventh International Conference on Information Visualization*, London (2003) 331-337.

[Gri02] Grinstein, G., Hoffman, P., Pickett, R., Laskowski, S., Benchmark Development for the Evaluation of Visualization for Data Mining, in Fayyad, U., Grinstein, G., Wierse, A. (Eds.) *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann, San Francisco (2002) 129-176.

[Gri06] Grinstein, G., O'Connell, T., Laskowski, S., Plaisant, C., Scholtz, J., Whiting, M., VAST 2006 Contest: A tale of Alderwood, *Proc. of IEEE Visual Analytics Science and Technology conference* (2006) 215-216

[Guo06] Guo, D., Chen, J., MacEachren, A. M., & Liao, K. 2006, A Visual Inquiry System for Space-Time and Multivariate Patterns (VIS-STAMP). IEEE Transactions on Visualization and Computer Graphics, 12(6), 1461-1474.

[Ham04] Frank van Ham, Technische Universiteit Eindhoven Contest Submission, *Poster Compendium of IEEE Information Visualization* (2004)

[Hof05] Heike Hofmann, Hadley Wickham, Dianne Cook, Junjie Sun, Christian Röttger, Boom and Bust of Technology Companies at the Turn of the 21st Century, *Poster Compendium of IEEE Information Visualization* (2005)

[Hon03] Hong, J. Y., D'Andries, J., Richman, M., Westfall, M., Zoomology: Comparing Two Large Hierarchical Trees, in *Poster Compendium of IEEE Information Visualization* (2003)

[Hos05] Bettina Hoser, Michael Blume, Jan Schröder, and Markus Franke, CPS- Company Positioning System: Visualizing the Economic Environment, *Poster Compendium of IEEE Information Visualization* (2005)

[Hsu04] Hsu Tzu-Wei, Lee Inman Farabaugh, Dave McColgin, Kevin Stamper, MonkEllipse, *Poster Compendium of IEEE Information Visualization* (2004)

[Inf04] Fekete, J.-D., Grinstein, G. and Plaisant, C., InfoVis 2004 Contest, www.cs.umd.edu/hcil/iv04contest

[Inf05] Grinstein, G., U. Cvek, M. Derthick, M. Trutschl, IEEE InfoVis 2005 Contest, Technology Data in the US, http://ivpr.cs.uml.edu/infovis05

[Inf06] InfoVis 2006 Contest http://sun.cs.lsus.edu/iv06/

[Kei04] Keim, D., Christian Panse, Mike Sips, Joern Schneidewind, Helmut Barro, University of Konstanz Contest Submission, *Poster Compendium of IEEE Information Visualization* (2004)

[Kei95] Keim, D., Bergeron, R. D., Pickett, R., Test datasets for evaluating data visualization techniques. In Grinstein, G., Levkowitz, H. , *Perceptual Issues in Visualization, Springer,* Berlin, (1995) 9-22

[Kom04] Komlodi, A., Sears, A., Stanziola, E., InformationVisualization Evaluation Review, *ISRC Tech. Report, Dept. of Information Systems*, UMBC. UMBC-ISRC-2004-1 http://www.research.umbc.edu/~komlodi/IV_eval (2004).

[Lee04] Lee Bongshin, Mary Czerwinski, George Robertson, Benjamin B. Bederson, PaperLens, *Poster Compendium of IEEE Information Visualization* (2004)

[Lin04] Lin Xia, Jan Buzydlowski, Howard D. White, Associative Information Visualizer, *Poster Compendium of IEEE Information Visualization* (2004)

[Las05] Laskowski, S., Plaisant, C., Evaluation Methodologies for Visual Analytics (section 6.1, in [Tho05] Thomas, J., Cook, K. (Eds.) *Illuminating the Path, the Research and Development Agenda for Visual Analytics*, IEEE Press (2005) 150-157

[Mac86] Mackinlay, J., Automating the design of graphical presentations of relational information, *ACM Trans. on Graphics*, 5, 2 (1986) 110, 141

[Min06] Minghim, R., Paulovich, F., Lopes A., Content-based Text Mapping using Multi-dimensional Projections for Exploration of Document Collections, *Proc. of Visualization and Data Analysis* (2006)

[MLR06] University of California Irvine Machine Learning Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html

[Mor03] Morse, D. R., Ytow, N., Roberts, D. McL., Sato, A., Comparison of Multiple Taxonomic Hierarchies Using TaxoNote, in *Poster Compendium of IEEE Information Visualization* (2003)

[Mul97] Mullet, K., Fry, C., Schiano, D., On your marks, get set, browse! (the great CHI'97 Browse Off), Panel description in *ACM CHI'97 extended abstracts*, ACM, New York, 113-114 (1997)

[Mun03a] TreeJuxtaposer, : InfoVis03 Contest Entry, James Slack, Tamara Munzner, and Francois Guimbretiere, Poster Compendium of IEEE Information Visualization (2003)

[Mun03b] Munzner, T., Guimbretière, F., Tasiran, S., Zhang, L. and Zhou, Y., TreeJuxtaposer: Scalable tree comparison using Focus+Context with guaranteed visibility. *ACM Transactions on Graphics*, SIGGRAPH 03 (2003) 453-462

[Nai01] North American Industry Classification System, www.census.gov/epcd/www/naics.html

[NYT06] New York Times – Election 2004 http://www.nytimes.com/packages/html/politics/2004_ELECTIONGUIDE_GRAPHIC/?oref=login (retrieved June 2005)

[Pal00] Pallett, D., Garofolo, J., Fiscus, J., Measurement in support of research accomplishments, *Communications of the ACM*, 43, 2 (2000) 75-79

[Pla02] Plaisant, C., Grosjean, J., and Bederson, B. B., SpaceTree: Supporting exploration in large node-link tree: design evolution and empirical evaluation, *IEEE Symposium on Information Visualization* (2002), 57-64.

[Pla04] Plaisant, C. The Challenge of Information Visualization Evaluation, in *Proceedings of the working conference on Advanced Visual Interfaces* (AVI 2004), pp. 109—116, Gallipoli, Italy, ACM Press, 2004.

[Sar04] Saraiya, P., North, C., Duca, K., An evaluation of microarray visualization tools for biological insight, *Proc. of IEEE Symposium on Information Visualization* (2004) 1-8

[Sch02] J. Scholtz, L. Arnstein, M. Kim, T. Kindberg, and S. Consolvo, User-Centered Evaluations of Ubicomp Applications, *Intel Corporation IRS-TR-02-006*, May 2002 2002.

[Sch05]Scholtz, J., Steves, M.P., A Framework for Evaluating Collaborative Systems in the Real World, to appear in *Proc. Hawaii International Conference on System Sciences*, 2005

[She03] Sheth, N., Börner, K., Baumgartner, J., Mane, K., Wernert, E., Treemap, Radial Tree, and 3D Tree Visualizations, in *Poster Compendium of IEEE Information Visualization* (2003)

[Shn06] Strategies for Evaluating Information Visualization Tools: Multidimensional In-depth Long-term Case Studies, Shneiderman, B., Plaisant, C., *Proc. of BELIV'06, BEyond time and errors: novel evaLuation methods for Information Visualization, a workshop of the AVI 2006 International Working Conference*, ACM (2006) 38-43

[SMo06] Smart Money Map of the Market www.smartmoney.com (retrieved June 2005)

[Spen00] Spenke, M., Beilken, C., InfoZoom - Analysing Formula One racing results with an interactive data mining and visualization tool, in Ebecken, N. *Data mining II*, (2000), 455–464

[Teo04] Soon Tee Teoh, Kwan-Liu Ma, One-For-All - University of California, Davis Contest Submission, *Poster Compendium of IEEE Information Visualization* (2004)

[Tra00] Trafton, J., Tsui, T., Miyamoto, R.; Ballas, J., Raymond, P., Turning pictures into numbers: extracting and generating information from complex visualizations. *International Journal of Human Computer Studies*, 53, 5 (2000), 827-850.

[Tho05] Thomas, J. and Cook, K. (Eds.) *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE CS Press (2005), http://nvac.pnl.gov/agenda.stm

[TRE06] Text REtrieval Conference (TREC), http://trec.nist.gov/

[Tym04] Jaroslav Tyman, Grant P. Gruetzmacher, John Stasko, InfoVisExplorer, *Poster Compendium of IEEE Information Visualization* (2004)

[Unw06] Unwin, A.R., Theus, M., Hofmann, H., Graphics of Large Datasets, Springer: New York (2006)

[VAS06] VAST 2006 Contest: www.cs.umd.edu/hcil/VASTcontest06

[Voo00] Voorhees, E., Harman, D., Overview of the sixth Text Retrieval Conference (TREC-6), *Information Processing and Management*, 36 (2000) 3-35

[Wei04] Weimao Ke, Katy Borner, Lalitha Viswanath, Indiana University Contest Submission, *Poster Compendium of IEEE Information Visualization* (2004)

[Whi06] Whiting, M. A., Cowley, W., Haack, J., Love, D., Tratz, S., Varley, C., Wiessner, K.; Threat stream data generator: creating the known unknowns for test and evaluation of visual analytics tools, *Proceedings of the 2006 Conference on Beyond Time and Errors: Novel Evaluation Methods For information Visualization* BELIV '06. ACM Press, New York, NY (2006) 1-3

[Won04] Wong Pak Chung, Beth Hetzler, Christian Posse, Mark Whiting, Sue Havre, Nick Cramer, Anuj Shah, Mudita Singhal, Alan Turner, Jim Thomas, IN-SPIRE, *Poster Compendium of IEEE Information Visualization* (2004)

[Zei05] Zeis Annerose, Sergej Potapov, Martin Theus, Antony Unwin Analyzing Company Data with Interactive Statistical Graphics, *Poster Compendium of IEEE Information Visualization* (2005)

Dr. Catherine Plaisant is Associate Research Scientist at the Human-Computer Interaction Laboratory of the University of Maryland Institute for Advanced Computer Studies. She earned a Doctorat d'Ingénieur degree in France in 1982 and joined HCIL in 1987. She enjoys most working with multidisciplinary teams on designing and evaluating new interface technologies that are useable and useful. She has written over 90 refereed technical publications on the subjects of information visualization, evaluation methods, digital libraries, universal access, image browsing, input devices, online help, etc. She co-authored with Ben Shneiderman the 4th Edition of Designing the User Interface.

Dr. Jean-Daniel Fekete is a Senior Research Scientist (DR2) at INRIA, one of the leading French National Research Centers, in Orsay, south of Paris. He leads the AVIZ Project since 2007, focusing on data analysis and visualization research. The AVIZ group is located in and collaborates with the Computer Science Department (the LRI) at the Universite Paris-Sud. AVIZ is studying multi-scale analysis and visualization of large datasets, combining machine learning approaches with information visualization and multi-scale interaction techniques to help analysts explore and understand massive data. Jean-Daniel's research topics include network visualization, evaluation of information visualization systems, toolkits for user interfaces and information visualization. His research is applied in several fields such as Biology, History, Sociology, Digital Libraries and Business Intelligence.

Georges Grinstein is Professor of Computer Science at the University of Massachusetts Lowell, Head of its Bioinformatics and Cheminformatics Program, Co-director of its Institute for Visualization and Perception Research, and of its Center for Biomolecular and Medical Informatics. His research interests include computer graphics, visualization, data mining, virtual environments, and user interfaces with the emphasis on the modeling, visualization, and analysis of complex information systems, most often biomedical in nature. He received his Ph.D. in Mathematics from the University of Rochester. He has over 30 years in academia with extensive private consulting, over 100 research grants, products in use nationally and internationally, several patents, and numerous publications. He has been on the editorial boards of several journals in Computer Graphics and Data Mining, a member of ANSI and ISO, a NATO Expert, and a technology consultant for various government agencies.